EOARD Special Contract
SPC - 95 - 4016

DEVELOPMENT OF COMPUTER METHODS FOR PREDICTION OF NEW MATERIALS HAVING PREDEFINED PROPERTIES

(final report)

Principal Investigators :

Professor Victor P. Gladun
Institute of Applied Informatics,
Kiev, Ukraine

&

Dr. Nellya D. Vaschenko
Institute of Applied Informatics,
Kiev, Ukraine

Members of Research Team :

Dr. Nadya N. Kiselyova
Baikov Institute of Metallurgy,
Moscow, Russia

Dr. A.L. Javorsky
Institute of Applied Informatics,
Kiev, Ukraine

August, 1995

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE August 1995	3. REPORT TYPE AND DATES COVERED Final Report

4. TITLE AND SUBTITLE

Development of Computer Methods for Prediction of New Materials Having Predefined Properties

5. FUNDING NUMBERS

F6170895W0220

6. AUTHOR(S)

Dr. Victor Gladun

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

V.M.Glushkov Institute of Cybernetics, National Academy Sciences Ukraine
Prospect Akademika Glushkova, 40
Kiev 252022
Ukraine

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0200

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

SPC 95-4016

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

A

13. ABSTRACT (Maximum 200 words)

This report results from a contract tasking V.M.Glushkov Institute of Cybernetics, National Academy Sciences Ukraine as follows: a new approach to prediction of new material having predefined properties that is based on the method of artificial intelligence.

14. SUBJECT TERMS

EOARD

15. NUMBER OF PAGES

49

16. PRICE CODE

N/A

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19, SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL

ABSTRACT

One of the trends involving the design of materials is the use of pattern-directed computational methods in the prediction of new materials from known compounds having the same or similar property values. One such computational method for discovering and predicting the properties of new materials, referred to as 'pyramidal networks', is to be discussed in this report. Pyramidal networks allow us to overcome difficulties connected with processing large data sets and extend previous classification methods by automating: 1) the formulation of generalized classes, 2) characterizing them by logical expressions, and 3) enabling continuous incremental updating of the classes and expressions. Methods of knowledge discovery and prediction involving the use of pyramidal networks are realized in a software application called CONFOR, which is further desribed in the appendix.

SUBJECT TERMS (Key Words)

Material prediction, knowledge discovery, concept formation, pyramidal network.

NOTICE

All inserted text, in boldface, has been added by Drs LeClair and/or Jackson to clarify the intent of the authors and to enhance readability of the report.

FOREWORD

This report was prepared by the above identified research team under EOARD Special Contract SPC-95-4016. This is the final report of the contract. We thank Dr(s) Steven R. Le Clair and Allen G. Jackson, researchers of the Materials Directorate, Wright Laboratory, for close collaboration in this work. We also thank Dr Osama El-Bayoumi (EOARD) for his support in the administration of the contract.

1. INTRODUCTION

1.1. Goals and Significance

The process of determining, *a priori* , the properties of new material compounds is divided into two steps:

1. Formation of concepts corresponding to classes of materials;
2. Use of the created concepts for prediction of new materials.

The first step is accomplished by forming a concept via the processing and generalizing over the attributes of a large set of examples; while the second step involves the application of a learned classification rule to predict the membership of a new material. Careful and rigorous exercise of the first step enables more accurate predictions. In general, this step involves the analysis of data containing descriptions of various compounds and/or processing parameters.

When designing new materials from patterns formed from knowledge of constituent elements or similar materials, at least two different situations are possible: 1) many similar materials having a complete set of predefined properties are known and available; 2) materials for which the complete set of predefined properties are unknown or there are only a few **[instances]** to establish any concepts.

Formulations of the problem and the methods of concept formation in these situations are quite different. In the first case, the problem of concept formation is transformed into the problem of inductive learning; **[generating some generalization from a specific set of instances].** In the second case, the prediction is an inference on the basis of analogy **[some interpolation or extrapolation from available data is required].** The research objective is to focus on the first case, and, more specifically, to enhance computational methods for new material prediction where there exist similar materials with a given set of properties.

1.2. Terminology and Problem

Prediction of new materials is based on knowledge characterizing conditions when similar materials were actually synthesized as well as cases when attempts to create such materials were unsuccessful.

> **[It is important to note that the authors are carefully establishing that both types of instances (i.e., successes and failures) are very useful and, in fact, his method referred to below as pyramidal networks makes specific use of each type of instance.]**

Knowledge is generalized information reflecting experience of observations and problem solving. By its nature knowledge is implicit information, in contrast to data which represents explicit information. Available methodology of knowledge formation is developed in the limits of such scientific trends as machine learning, inductive inference, dependencies discovery, concept formation, inference on the basis of analogy [1-9]. Lately some new terms have appeared such as "knowledge discovery", and "knowledge mining" that usually designate methods of revealing dependencies among data in data bases [10].

Knowledge is formed by revealing regularities that are inherent in data. There exist three types of regularities, namely:

o characterizing dependencies between objects;

o characterizing sequences of objects;

o characterizing sets (classes) of objects.

[Note that the authors use the word 'characterizing' to preface each type of regularity (dependency, sequence, and set). The inference here is to be expanded upon further below. The significance of their use of the word 'characterizing' is to establish that a regularity may be represented by a subset of the attribute description of the objects.]

Revealing regularities of various types is accomplished by quite different methods. Having as a goal new materials prediction, we face the problem of revealing knowledge characterizing sets or classes of objects. In this case the word "object" designates any element and/or compound of the material world that can be represented by a set of attributes. In particular an object can be a situation as well as a realization of some process or phenomenon. The word "attribute" means everything that characterizes the object and that can be used in such operations as selection, recognition, comparison, and so on. Knowledge about a class of objects has a form of a concept.

A concept is a generalized model of some class of objects that is used for recognizing and generating models of specific elements of this class. Relative to a word, a concept is its sense, i.e., information about realities (denotates) that are designated by the word. The set of denotates is the volume of the concept. A concept is a lexical rule defining application of a word.

A concept is an important tool for problem solving and formation of new knowledge. Concepts are necessary in processes of classification, generalization, and organization. In these processes two main functions of a concept are realized, namely: recognition and generation of models of its denotates. Process of recognition has become an object of research and automation long ago, while the generation of models is still an unresearched problem. Models generation takes an important place in creative activity. It lies at the heart of engineering design.

[Note that the authors appear to be stating that class recognition via characterizing set of attributes is distinguished from models of relationships between attributes in the class.]

Attributes characterizing a concept are divided into separating and unifying ones. Separating attributes are those attributes that do not occur or occur seldom outside the concept volume. These attributes are the most useful for implementation of the recognition function. Unifying attributes belong to all denotates or to most of them, but they can be found outside the concept volume as well. Unifying attributes are necessary for generation of object models.

[It is important to note that the authors are careful to establish that, given the application of some method to characterize a class, a concept is effectively a set of separating **attributes which the user has *a priori* determined to be those attributes which constitute the class. For example, assume for the moment that the universe of all attributes is the twenty-six (26) letters of the alphabet: A, B, C, . . . X, Y, and Z. Now also assume based on my experiences that I have recognized a pattern, a set of occurrences (object instances), defined as class (I), of material to be represented by "**unifying**" attributes A, C, and E, and a second class (II) of material to be represented by "**unifying**" attributes C, D and F. Now among these classes of materials, a subset called "**separating**" attributes A and E constitutes a model distinguishing the class I from II.]**

Volumes of concepts that were not defined artificially as a rule do not have precise boundaries.

[The authors are simply acknowledging that separating attributes are based upon human judgement.]

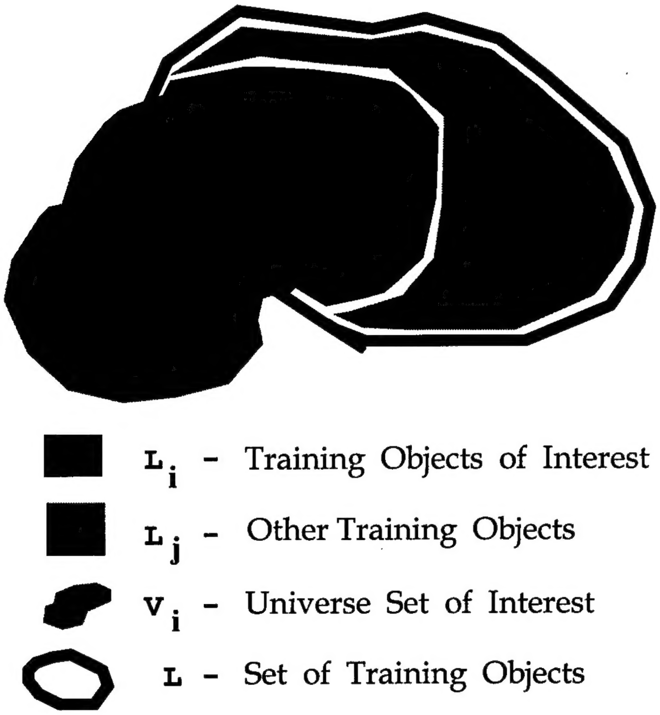
Any system of concepts does not reflect the variety of the real world. So accurate identification of objects sometimes becomes quite difficult or even impossible. Therefore, volumes of many concepts can be considered as fuzzy sets, i.e., as those sets that have a membership function indicating the degree of belonging of objects to the set. The membership function has, of course, a subjective nature. Concepts corresponding to fuzzy sets are referred to as fuzzy concepts.

The problem of inductive concept formation for n disjoint sets (classes) V_1, V_2,...,V_n is defined as follows. Let L be a set of objects used as a training set. Relationships $L \cap V_i$ _ $\varnothing$ and $V_i \not\subseteq L$ (i=1,2,...,n) are assumed. All elements of set L are represented by attributive descriptions. Each object $l \in L$ has an indication of a class it belongs to. It is necessary by analyzing L to form n concepts Q_1, Q_2,...Q_n with volumes V_1, V_2... V_n providing correct recognition of all objects $l \in L$.

[The following is provided by Dr LeClair to further detail the above discussion by the authors on the forming of a concept. It should be noted that the disjoint sets, i.e., classes, are assumed to be defined by the user in all cases.]

[Using the depiction below, let L be a set of training objects l such that L (intersected with) V_i is not equal to empty set (L containes objects from each class V_i) and V_i are not wholly included in L, where (i = 1, 2,..., n). .

All elements of set L are represented by attributive descriptions. Each object l (is an element of) L is also a candidate object l (element of) V_i (i = 1, 2,..., n). V_i is an object set (class) that we want to investigate, yet we know only some of the objects l of the set V_i are within L, i.e., L_i represents the subset of training objects V_i which are contained in L. When forming a concept we analyze L_i (and all other objects in training set), but in analyzing L_i, we build a concept for V_i. The concept resulting from analysis may not recognize correctly all remaining objects of a set V_i (not included in L_i) but the goal is to build a concept which is as accurate as is possible in predicting new objects of the universe set V_i.]


L_i - Training Objects of Interest
L_j - Other Training Objects
V_i - Universe Set of Interest
L - Set of Training Objects

The formed concepts afterwards are used for classification of new objects, diagnosis, and prediction. The concept formed on the basis of a training set in the general case is some approximation of the real concept. Closeness of these concepts depends on representativeness of the training set (how completely it reflects <u>pecularities</u> of the corresponding class).

[It is again important to note what the authors are saying in their reference to 'pecularities'. In effect, a concept is a conjunction of attributes and those attributes are, in fact, exemplar attributes of the

class. If one or more of the exemplar attributes are missing, because the training set was not representative enough of the class, then the attributes which comprise the concept will lack the ability to predict new class members.]

The problem of inductive concept formation is quite similar to the problem of pattern recognition learning. In both cases, learning results in the forming of a model of some class of objects. For concept formation, this model (concept) must satisfy additional strong requirements: it must provide not only recognition but also generation of object models. Therefore, it must reflect attributive, structural, and logical characteristics of objects. For example, when forming a concept corresponding to class V_i, objects of the training set belonging to V_i are considered examples of class V_i (positive objects); other objects are referred to as negative objects.

Central to the methods of concept formation is the process of selection of attribute combinations characterizing classes of positive and negative objects. In this process it is quite important to minimize examinations of large volumes of data.

[Two points of importance to note here: 1) concepts may be characterized by both positive and negative objects, 2) given a manual method of concept formation - when establishing a training set it is important to limit the size of the training set. The reason becomes apparent in practice, but to be succinct here - a large training set will confound the method, making it difficult to separate the "signal from the noise".]

2. CONCEPT FORMATION AND PREDICTION IN PYRAMIDAL NETWORK

2.1. Pyramidal Networks

2.1.1. Definitions

Let K be a combination of attributes selected at some stage of concept formation; l be attributive description of object l $\in$ L.

Set $E = \{K : |K \cap l\,| > 1\}$ is called encirclement of object l.

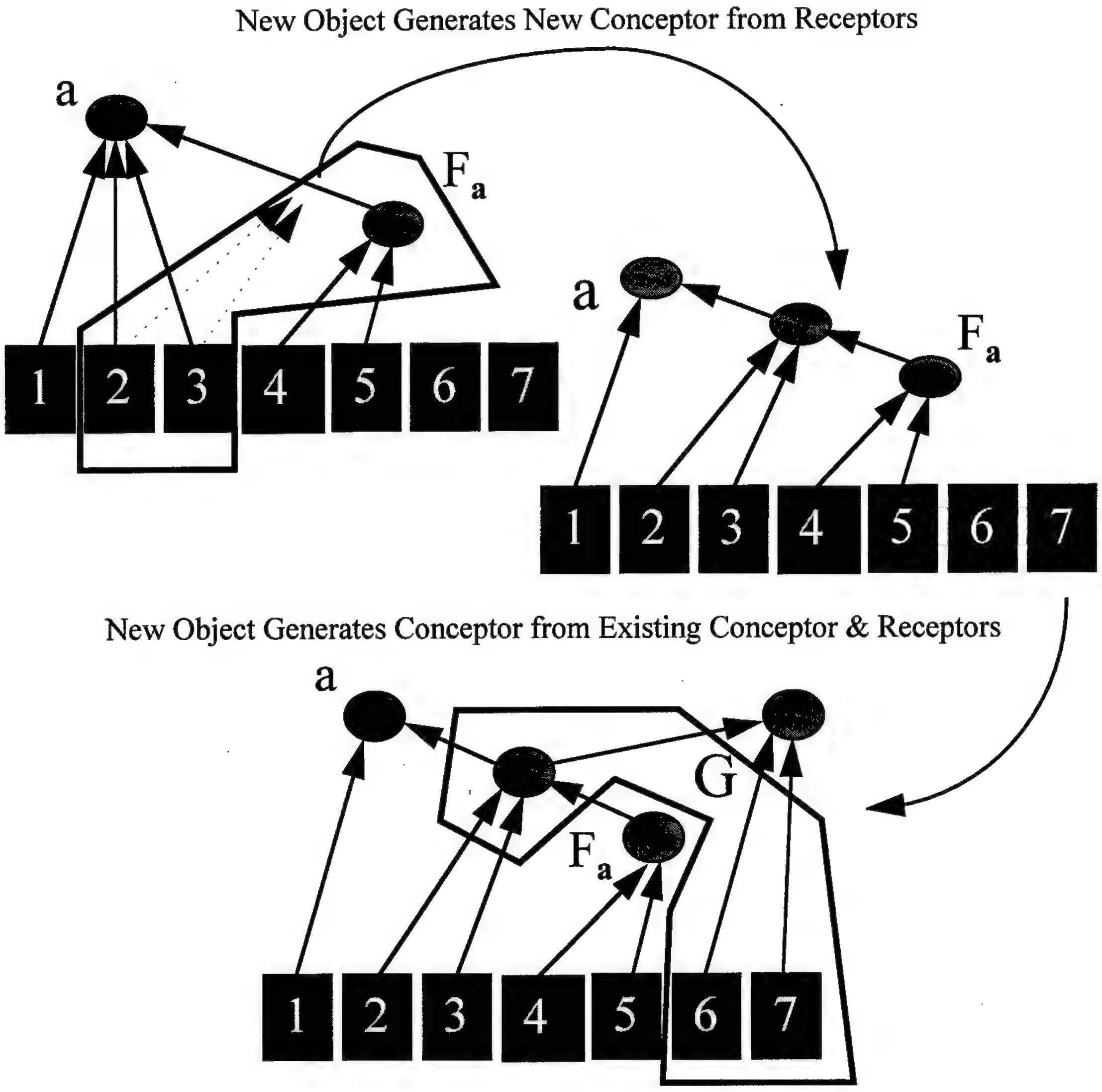
A method of knowledge discovery is referred to as a 'local statistical method' if all search operations, when object analyzing, are realized in the limits of this object encirclement. With such methods, the selection of new attribute combinations is dependent on previous combinations.

[This simply means that <u>more</u> than one (1) receptor is necessary to associate an object with a concept, and concepts evolve via the incremental (with the evaluation of each new object of the class) construction of attribute conjunctions.]

A typical example of local statistical methods is the method of concept formation in pyramidal networks [4-6,11-16,18-20]. Application of pyramidal networks overcomes **[reduces]** difficulties connected with examination of large data volumes. Another advantage of the method consists in the formation of concepts with exclusive attributes, i.e., with attributes that do not belong to positive objects. Such type of knowledge usually provides a more accurate prediction or diagnosis. Numerous applications of the method, in particular for predicting new materials [17], demonstrate its efficiency.

A pyramidal network is an acyclic graph having <u>no</u> vertices with a single entering arc. Examples of pyramidal networks are given in Figures 1-3. Vertices having no entering arcs are referred to as receptors. Other vertices are referred to as conceptors. As depicted in Figure 1, receptors are used to form conceptors. The generation of conceptors is incremental with one object at a time, wherein, if the existing conceptors do not satisfy the attribute description of the new object, a new conceptor is formed, either from receptors or the combination of conceptors and receptors.

[Figure 1 illustrates the evolving nature of the PN and begins to capture the essence of concept formation referred to as encirclement of an object. As each new positive object (object of the class of interest) is considered for representation by the PN, if no existing conceptor is appropriate, two (2) or more receptors of the object are used to form a new conceptor to characterize the object. The new conceptor may also contain one of the existing conceptors together with one or more of the receptors.



New Object Generates New Conceptor from Receptors

New Object Generates Conceptor from Existing Conceptor & Receptors

Figure 1. Creating a Pyramidal Network

A subgraph of the pyramidal network including vertex **a** and all vertices from which there are paths to the vertex **a** is the pyramid of vertex **a**. The set of vertices contained in the pyramid of vertex **a** is referred to as subset of vertex **a**. The superset of vertex a is defined as a set of vertices towards which there are paths from vertex **a**. For example, the **O**-subset **[1, 2, 3, & F$_a$]** and **O**-superset **[not defined in Figure 1]** of vertex **a** consist of vertices that are connected with vertex **a** directly.

2.1.2. Algorithm of Network Formation

Depending on the applications, receptors of a pyramidal network correspond to attribute values and, therein, can be numbers, numerical intervals, as well as names of relations, properties, states, actions, objects, and classes of objects.

New information **[objects]** entered into the network is represented as a set of attributes values. Corresponding receptors **[i.e., receptors which correspond to attributes describing the object]** switch to a state of excitation. The process of excitation is propagated in the network. A conceptor switches into the state of excitation if all vertices of its **O**-subset are excited. Receptors and conceptors retain their state of excitation during execution of all operations connected with network building.

Assume a new object **[with attributes 2, 3, 4, 5, 6, & 7]** enters the network of Figure 1. Let $\mathbf{F_a}$ be a subset of excited vertices of the **O**-subset of vertex **a**, and let **G** be a complete set of excited vertices such that no other excited vertices exist in their supersets. New vertices are added to a network according to the following two (2) rules:

> **A1**. If vertex **a** is not excited and $|\mathbf{F_a}| > 1$, the arcs connecting vertices of set $\mathbf{F_a}$ with vertex **a** are eliminated and a new conceptor is added to the network that is connected with vertices of set $\mathbf{F_a}$ by entering arcs and with vertex **a** by outgoing arcs. The new vertex is in the state of excitation.

Rule A1 is illustrated in Figure 1 (Networks I and II). Network II appears after excitation of receptors 2,3, 4, and 5 in situation I.

After introducing new vertices into all sections of the network where the condition of rule A1 is satisfied, rule A2 is performed.

> **A2**. If $|\mathbf{G}|>1$, a new conceptor is added to a network, which is connected with all vertices of set **G** by entering arcs. The new vertex is in the state of excitation. Rule A2 is illustrated in Figure 1 (Networks II and III). Network III appears after excitation in situation II of receptors 2,3,4,5,6, and 7.

In the case when sets of attribute values are descriptions of some objects (e.g., materials) conceptors formed by rule A1 represent intersections of some of these descriptions and conceptors formed by rule A2 correspond to complete descriptions.

2.1.3. Properties of Pyramidal

Pyramidal networks are convenient for execution of various operations of associative search. For example, it is possible to select all objects that include some combination of attribute values by tracing outgoing paths from the network vertex that corresponds to this combination. To select all objects having common attribute values with given objects it is necessary to trace outgoing paths from vertices of its pyramid. The algorithm of network formation automatically identifies associative proximity of objects having common combinations of attribute values. All processes connected with the processing of an object description are localized in a relatively small part of a network - in the object pyramid.

An important property of pyramidal networks is their hierarchical nature which makes it possible to represent the structure of compound objects in a natural way. For example, conceptors of a network correspond to conjunctive combinations of attributes **[receptors]** which describe objects. If an object pyramid includes an excited vertex, then this object is connected with the class represented by this vertex. Thus, during network construction classification of objects is carried out.

Information about objects and classes of objects is represented by groups of memory elements distributed over the whole network. Certainly, the advantages of pyramidal networks can be seen best in their physical realization which permits parallel propagation of signals through the network. An important property of the network as a means for information storage is the fact that the possibility of parallel propagation of signals is combined in the network with the possibility of parallel reception of signals from sensors of a system.

Pyramidal networks in combination with methods of their construction has become a convenient tool for mass data representation in different applications.

2.2. Algorithm of Concept Formation

Let there be a growing semantic network representing a training set L. Such a network can be built by application of the algorithms of network formation for each $l \in L$. To form concepts $Q_1, Q_2, ..., Q_i$ it is necessary to examine sequentially pyramids of all objects of the training set. In the process of this examination special vertices are chosen that should be used for recognition **[note, not model generation]** of objects from concept volumes. They are referred to as check vertices.

To choose check vertices two characteristics of the network vertices are used: m_i and k, where m_i is the number of objects from the volume of concept Q_i whose pyramids include the given vertex **[assume a class of objects; for each conceptor in the network, m establishes the number of objects it represents in that class]**, and k is the number of receptors in the pyramid of this vertex **[again assume a class of objects, k establishes the number of receptors it represents]**. For all receptors, k = 1. When examining a pyramid the transformations described by the rules given below are performed.

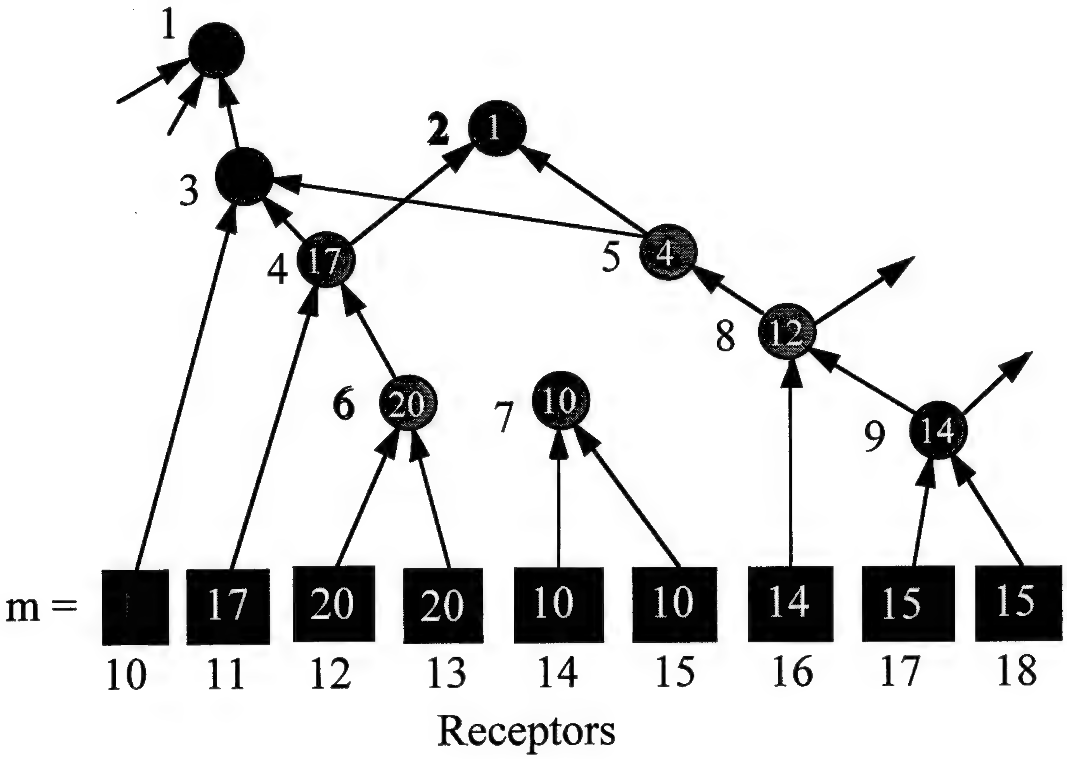


Figure 2. Determining Check Vertices in a Pyramidal Network

B1. If in the pyramid of some object belonging to the volume of concept Q_i the vertex having the largest k_i **[receptors]** among all vertices with the largest m_i **[objects]** is <u>not</u> a check vertex of concept Q_i, then it is marked as a check vertex of the concept Q_i.

The action of rule B1 is illustrated in Figure 2 as follows: the excitation of the pyramid of vertex 2 will result in choosing vertex 6 as a check vertex, because it has the largest k **[it has two receptors, whereas vertices 12 and 13 are receptors, i.e., they have only one receptor]** of all vertices having the largest m_i **[they each have 20 objects]**.

B2. If the pyramid of some object from the volume of concept Q_i has check vertices of other concepts which do <u>not</u> contain <u>in their supersets</u> the excited check vertices of concept Q_i, for each of these supersets the vertex having the largest k among all excited vertices with the largest m_i is marked as a check vertex of the concept Q_i.

[The check vertex for each object of a class (I) must <u>not</u> have a check vertex for some other class (II) in its superset; if it does, then the check vertex for the class (I) must be changed.]

According to this rule, the excitation of the pyramid of vertex 2 (Figure 3.I) results in choosing vertex 5 **[instead of vertex 6]** as a check vertex of concept Q (Figure 3.II).

With the help of check vertices the most typical (having the largest m_i) **[associated with the largest number of objects in the class]** combination of attributes are established as concepts. **[for recognition of the class]** For example, choosing vertex 8 (Figure 3.I) as a check vertex means receptors 16, 17, and 18 constitute the most typical combination of attributes associated with a particular class of objects. If at least one new check vertex appears when examining new objects of the training set, a new examination of the complete training set is carried out. The algorithm operation is completed if during the examination of the training set no new **[additional]** check vertex appears. **[as a consequence of rule B2]** A concept that arises as a result of algorithm execution is represented by some collection of check vertices.

> **[a concept is a collection of vertices because a concept may be comprised of multiple positive and/or negative check nodes.]**

By analyzing the network it is possible to construct a description of the concept in the form of a logical expression [12-16].

2.3. Prediction of New Materials

To predict if a new material will have the desired property, its attribute description should be classified on the basis of knowledge (a concept) of **[contained in]** a class of materials with the same properties. After learning **[training]** of pyramidal networks **[has been accomplished]** the following classification rule is used.

An object enters into the volume of concept Q if in its pyramid there are check vertices of concept Q and there are no check vertices of any other concepts in its superset. Thus, new material with predefined properties can be classified by calculating the value of the logical expression **[i.e., a collection of check vertices]** that represents the corresponding concept.

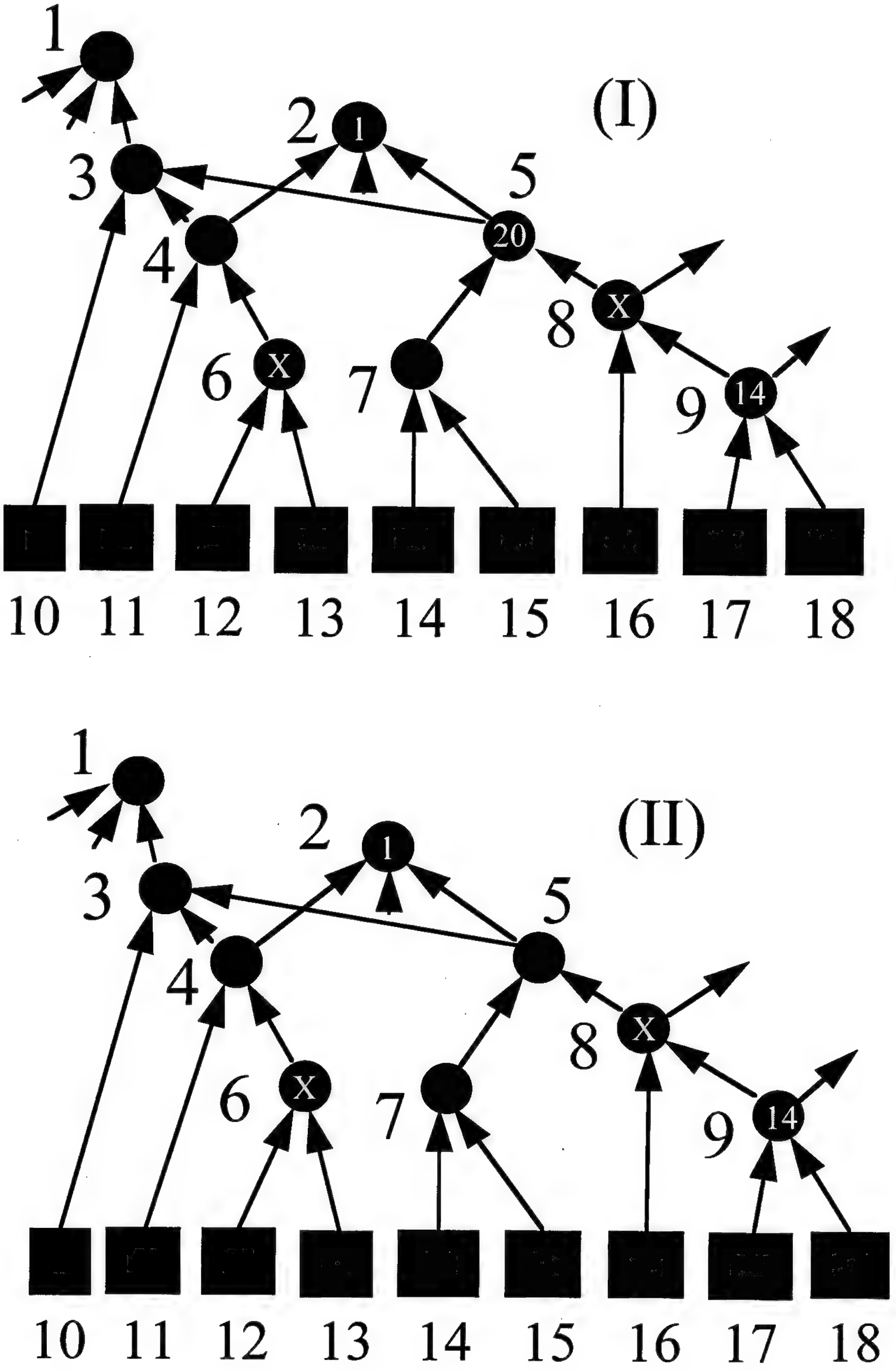


Figure 3. A Pyramidal Network

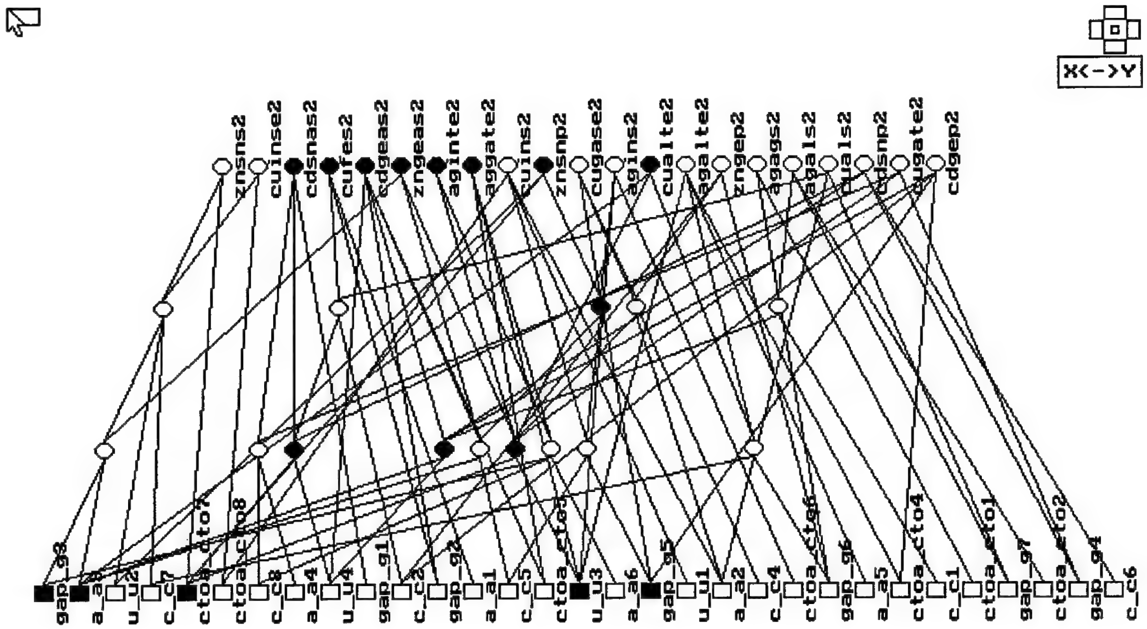


Figure 4. A pyramidal net for a data set of chalcopyrite compounds. The red represent check vertices of one class, the blue those of the second.

[An example of a pyramidal net with check vertices indicated for two classes is given in Figure 4. This set of 26 chalcopyrite compounds contained five variables, each of whose values was arbitrarily binned to produce a small set of attribute values associated with each variable. The algorithm was applied to the data and the resulting net is that shown in the figure. By considering the check vertices, compounds that are characteristic of a class are easily identified. A new chalcopyrite compound can be quickly classified by comparison with these check nodes and their contents. As described, the check vertices represent logical conjunctions and disjunctions of attributes for the classes considered, and, therefore, are representative of these chalcopyrites. We note that the net will change considerably vis a vis the check vertices when the class definitions are changed.]

2.4. Some Fundamental Characteristics of the Method

It has been proved [5,12,15,16] that the algorithm of concept formation in pyramidal networks is convergent and provides formation of the concepts dividing a training set of any complexity.

Concepts are represented by collections of check vertices or by logical expressions. It is important that the algorithm provides the possibility to include in the concept exclusive **[negative]** attributes that do not belong to positive objects. As a result concepts formed by the algorithm have more perfect logical structure. It simulates a more accurate prediction or diagnosis. In logical expressions exclusive attributes are represented by variables with negation.

All search operations are limited to a comparatively small section of the network that includes the pyramid of the object and vertices directly connected with it.

To understand algorithms of such sort, sometimes geometric interpretation is useful. Geometric interpretation of concept formation in pyramidal networks is given in [5,12,15,16].

A vertex having in its pyramid k receptors is represented in an **n** dimensional space of attributes by an (**n-k**) dimensional plane. Multi-dimensional planes corresponding to check vertices of concept Q are called zones of concept Q. Each concept is represented in the attribute space by an area consisting of zones of this concept. This area approximates the area of positive objects distribution.

[The reference to 'n-k' dimensional plane is intended to convey that a concept will typically be comprised of a subset of receptors. The totality of receptors is 'n', and the subset representing a vertex is always 'n-k', where 'k' represents all receptors not connected to the vertex of interest. Thus, a zone of concept Q is the collection of receptors connected to a vertex. Border zones represent receptors which are connected to multiple vertices and thus multiple concepts.]

Building an approximating area for concept Q consists of two processes:

1) a rough covering area of positive objects distribution in the attribute space by zones of concept Q (rule B1);

2) a division of so-called "border" zones, i.e., zones that contain points representing objects of different classes (rule B2).

Mathematical models of pyramidal networks are published in [15,19].

3. SPECIFICS OF KNOWLEDGE DISCOVERY PROCESS

Quality of knowledge depends on:

1) effectiveness of an algorithm of knowledge discovery;
2) representativeness of a training set (first of all, how completely does it represent "border" objects, i.e., those objects that in an attribute space are represented by points near the surface separating positive and negative areas;
3) chosen attributes.

3.1. Choice of an Attribute Set

Taking into account the dual purpose of a concept (for classification and model generation) it is necessary to include both separative and unifying attributes.

Numerous applications of the system CONFOR [20] involving concept formation using pyramidal networks have resulted in methodologies for attribute set formation. The methodologies are based on an understanding of knowledge discovery as an interactive process, each step of which focuses on the perfection **[refinement]** of an attribute set, as well as the knowledge that has already been revealed. If the discovered knowledge for some reason is not satisfactory (for example, ineffective for classification, diagnostics, or prediction) a researcher changes the attribute set and repeats the process. It is important that the form of knowledge representation at the system output should be convenient for the formation of decisions relative to the necessary corrections of the attribute set.

For concept formation in pyramidal networks, discovered knowledge is represented with a logical expression containing the same designations of attributes that were used in descriptions of the training set. Each conjunction contained in the logical expression is followed by a number indicating how many times it occurs in object descriptions. Similarities in output and input representations are simplified using comparisons of discovered knowledge with object descriptions, and furthermore, result in the formation of ideas concerning improvements in the attribute set.

Changes **[refinements]** in the attribute set involve application of the following operations:

o excluding attributes belonging to both positive and negative objects,

o unifying correlating attributes, and

o introduction of new attributes.

Appropriate choice of the attribute set allows formation of concepts of the most simple, logical structure.

3.2. Discretization of Analog Values

Methods of knowledge discovery are usually applied to attributive descriptions in which each attribute has a finite number of values. Therefore, at the stage of description formation ranges of numerical values are divided into nonintersecting subintervals, each of which corresponds to one discrete value of an attribute. An operation of discretization (binning) is fulfilled on the basis of comparison of distributions of training set objects in scales of numerical values of attributes. Subintervals with the biggest density of object distribution, as well as subinterwals containing separating values of the same class, are choosen as discrete values of attributes.

[The subject of discretization is treated very briefly here. Discretization is simply the converting of a continous, piecewise continuous range of attribute values to intervals which can be represented as discrete ranges which greatly assist in pattern association.

It is important to note that discretization may be the single most important contributant to the formation of concepts of "simple, logical structure" as referenced above.]

3.3. Processing of Uncertainty in a Training Set

A training set may contain uncertainty of the following types:

1) absence of information about a value of an object attribute;
2) vagueness of information about a value of an object attribute;
3) coincidence of object descriptions belonging to different classes.

These uncertainties are usually encountered in using *a priori* estimations of trustworthiness of an attribute value of an object. For example, such estimation can be the probability of belonging of an attribute value to an object. When applying this approach the main difficulty is connected with the definition of initial probabilities.

In practice, initial probabilities are often given only on the basis of voluntary, subjective opinions. Therefore, it is better to omit doubtful attributes, believing that, for a representative training set, completeness of some object descriptions should not be allowed to unduly influence the quality of some discovered knowledge.

Coincidence of object descriptions of different classes can be caused by imperfection of the attribute set. It takes place also in the case of fuzzy concepts.

[Knowing, *a priori* the best, most descriptive, set of unifying and separating attributes which establish, via training of pyramidal network, a class is clearly a limitation of concept formation.]

In particular, coincidence of object descriptions can be a result of excessive generalization of an attribute. If coincidence of descriptions can not be removed by changing of an attribute set, it is necessary to exclude coincident descriptions from the training set or to use a non-deterministic method of concept formation in pyramidal networks [12]. In this non-deterministic method, objects belonging to the area of intersection of concept volumes are classified on the basis of Bayes rule.

3.4. Non-uniqueness of Representation of Discovered Regularities

Knowledge revealed on the basis of training set analysis can be represented by logical expressions. Usually a set of attribute descriptions can be represented by more than one logical expression. Methods of knowledge discovery may be compared on the basis of the formed logical expressions reflecting the quality of the revealed knowledge. **[- but the -]** Desire to form the best logical expression should be balanced by the understanding that the main factor influencing concept quality is representiveness of the training set.

The method of concept formation in pyramidal networks belongs to a class of methods where the formed logical expressions depend on the order in which training set objects are examined. In methods of this type, it is very important to conserve the quality of discovered knowledge when changing the order of a training set examination. Differences in logical expressions are not so important if the expressions are equally effective in application, i.e., classification, diagnostics, and prognostication.

Pyramidal networks, to a large degree, conserve knowledge quality, in spite of changes in the order of object examination, because at each step of concept formation the prior concept is corrected to provide accuracy of classification of the object under consideration. The effect of knowledge quality conservation in spite of changes in the order of training set object examination is a consequence of the algorithms used for adaptation or corrective actions to situations that appear at intermediate stages of training set processing.

4. CONCLUSION AND PROPOSALS

The merits of pyramidal networks for discovery of complicated regularities in large scale datasets are as follows:

- o use of a special data structure (pyramidal network) providing simplification of search operations;
- o knowledge representation using attributes of negative objects that provides discovery of more general regularities.

Success of knowledge discovery and prognostication depends on the selected attribute set, in particular, on the selection of interval attributes in numerical scales **[discretization]**. It is important to address the problem of discretization. In the process of further development of the CONFOR system it is necessary:

- o to replace existing data input facilities with an inter-application tabular (spreadsheet) input;
- o to add in the program "Search for combinations" the possibilities: 1) to output revealed combinations of attribute values in the order according to the number of their occurence in the training set; and 2) to output names of objects that contain the revealed combination;
- o to print the network picture;
- o to supply the system **[CONFOR]** with a mode of learning on the basis of additional data;
- o to organize an 'interrupt' capability in learning and recognition modes;
- o to output the logical expression together with the dictionary (for convenience in testing correspondence between bins and their names in a logical expression);
- o to form object descriptions using **[more natural language]** descriptions of their compounds.

It is quite important to create methods and program tools for designing technologies of new materials synthesis methods in cases where a training set is not sufficient. This is a very promising trend in the development of computer technologies of new materials design. It is also expedient **[of entrepreneurial interest]** to organize an international service for supplying scientific and industrial organizations with predictions of new materials synthesis.

REFERENCES

1. E.B.Hunt. *Concept Formation: An Information Processing Problem.* Wiley, New York, 1962.

2. M.M.Bongard. *Problems of Recognition* (in Russian). Moscow: Nauka, 1967, 320 p.

3. R.B.Banerji. *Theory of Problem Solving*. American Elsevier Publishing Company, New York, 1969.

4. V.P.Gladun. *Concept Formation by Growing Networks Learning* (in Russian). Kibernetika, No. 2, 1970, pp. 99-104.

5. V.P.Gladun. "Computer Description of Classes of Objects." Cybernetics, 824-832 (Trans. from Russian: Kibernetika, N 5, 1972, pp. 109-117; No. 6, 1972, pp. 28-36).

6. V.P.Gladun, N.D.Vaschenko. "Methods of Forming Concepts with Computer." Cybernetics, 295-301 (Trans. from Russian: Kibernetika, No. 2, 1975, pp. 107-112).

7. A.G.Arkadjev, A.L.Braverman. *Computer Learning to Classify Objects* (in Russian). Moscow: Nauka, 1971, 192 p.

8. V.Kodratoff and R.S.Michalski, editors. *Machine Learning, an Artificial Intelligence approach*, v.3. Morgan Kaufmann, San Mateo, California, 1990.

9. R.S.Michalski, J.G.Carbonelli and T.M.Mitchell, editors. *Machine Learning, an Artificial Intelligence Approach*, Vol. 1, Morgan Kaufmann, San Mateo, California, 1983.

10. G.Piatetsky-Shapiro and W.J.Frawley,editors. *Knowledge Discovery in Databases*. AAAI Press, Menlo Park, California, 1991.

11. N.D.Vaschenko. "Concept Formation in Semantic Network" (in Russian). Kibernetika, No. 2, 1983, pp. 101-107.

12. V.P.Gladun. *Heuristic Search in Complicated Environments* (in Russian). Kiev: Naukova Dumka, 1977, 166 p.

13. V.P.Gladun, Z.L.Rabinovich. "Formation of the World Model in Artificial Intelligence Systems." In: Machine Intelligence, 9, Ellis Herwood Ltd., Chichester, 1980, pp. 299-309.

14. V.P.Gladun. "Growing Semantic Networks. Computers and Artificial Intelligence," Bratislava, Nos. 1 & 5, 1986, pp.13-27.

15. V.P.Gladun. *Solution Planning* (in Russian). Kiev: Naukova Dumka, 1987, 168 p.

16. V.P.Gladun. *Processes of New Knowledge Formation* (in Russian). Sofia: SD Pedagog, 1994, 192 p.

17. E.M.Savitski, V.B.Gribulya, N.N.Kiselyova, et al. *Computer Aided Prognostication of Materials* (in Russian). Moscow: Nauka, 1990, 86 p.

18. V.P.Gladun, N.D.Vaschenko. “Adaptive Problem-Solving Systems.” Kybernetes, Vol. 9, 1980, pp. 181-188.

19. M.B.Burgin, V.P.Gladun. “Mathematical Foundations of Semantic Networks Theory.” In MFDBS 89, 2nd Symposium on Mathematical Fundamentals of Databases Systems, Visegrad, Hungary, June, 1989, Proceedings. Lecture Notes in Computer Science, 364, Springer Verlag, pp.117-135.

20. N.D.Vaschenko.CONFOR “Tools for Regularities Revelation and Analysis”. Upravljajuschie sistemi i mashini, 5/6, 1992, pp. 26-30.